Make Chicago "Safe" Again

Data Mining Project

Luis Veltze
University of Colorado at Booulder
Boulder, Colorado
luis.veltze@colorado.edu

Ryan Close
University of Colorado at Boulder
Boulder, Colorado
ryan.close@colorado.edu

Tyler Mooney
University of Colorado at Boulder
Boulder, Colorado
tyler.mooney@colorado.edu

Garrett Glissmann
University of Colorado at Boulder
Boulder, Colorado
garrett.glissmann@colorado.edu

ABSTRACT

The goal of this project is to find interesting trends using Chicago's crime reports spanning the years 2001 to early 2018. We will perform a longitudinal survey of crime trends. We hope to find geographical patterns as well as discover connections to other public health factors and socioeconomic indicators. For example, we could correlate crime with birth rates or access to affordable housing and grocery stores.

ACM Reference format:
Luis Veltze, Ryan Close, Tyler Mooney, and Garrett Glissmann. 2016. Make Chicago "Safe" Again. In *Proceedings of CSCI-4502,*
, Boulder (University of Colorado at Boulder), 5 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 PROBLEM STATEMENT

In our search to find a interesting dataset that would be a good candidate for data mining, we considered datasets about bitcoin and crime. We selected crime data because it was more intuitive to understand the attributes and ultimately mine correlations between similar datasets. Crime data is one of many subjects that is continuously examined and analyzed for trends and patterns. Of course, studying crime has an important social value: to understand how much crime is happening, how it compares in terms of time and geography, and possible reasons that crime rises or falls. The hope is that in understanding the patterns and causes of crime, it might be possible to bring about changes that can reduce crime. This is evident in policing today in terms of directed patrols to crime "hotspots" within cities. Using the crime data and other public data recorded by the city of Chicago, this team will search for trends in crime that perhaps could be used for social action.

University of Colorado at Boulder,

DOI: 10.1145/nnnnnnn.nnnnnnn

2 LITERATURE SURVEY

The existing content analyzed in this project often categorizes types of crimes into two groups: index and non-index crimes. Index crimes include murders, criminal sexual assaults, aggravated assaults/batteries, burglaries, thefts, robberies, arson, and motor vehicle thefts. Non-index crimes include all other crimes, such as vandalism, weapons violations, public peace violations, and others. Sources, such as the report by Andrew Papachristos describe the spatial trends of crime in Chicago as well as the changes in terms of frequency over time of index crimes and the overall crime rate. According to Papachristos' report, Chicago's crime rates are similar to those of other U.S. cities. Crime rose in the 1960s, reached a high point in the 1990s, and since then it has steadily declined (Papachristos 4). Additionally, the report mentions that "especially socially and economically disadvantaged communities continue to have stubbornly high levels of crime" (Papachristos 6). Given the data that is used in this project, the team should be able to see this second trend in the rates of crime by community area in Chicago. Other sources found online present visualizations detailing the prevalence of types of crimes, such as theft versus battery. They also present time series plots of crimes over 17 years. An additional point of study was the trends of a particular crime, such as sexual assault, over a single year (Laughlin; Mangipudi).

3 PROPOSED WORK

3.1 Data Cleaning:

We will need to clean all of the data sets that we plan on using. This will involve ensuring that there are no missing or na values in the data sets and dropping unnecessary columns or empty rows from the data set. We will need to ensure that the data is in the correct format for use with other tools.

3.2 Data Preprocessing:

To ensure that our data is easy to use with the variety of tools that we intend to use on this project, we will need to spend a fair amount of time on data preprocessing. We will begin by calculating various metrics on the data, such as average crime rate, change in crime rate, arrest rate, and so on. Calculating this meta-data will be useful when we are creating visualizations of the data. Because a lot of the data is text-based, we will explore different methods

of tokenization that will allow us to use the text-based data with various machine learning tools.

3.3 Data Integration:

We are considering using weather data and census data to help us find potential relationships between the data sets. Our hope is to see some sort of relationship between the temperature and the crime rate. We will also attempt to see if there is a relationship between the average income of a neighborhood and the type of crime that is typical to the area.

3.4 Visualization:

Visualizing our data will be integral to finding interesting patterns and correlations in the data set. We will be looking for interesting trends in the crime rate in Chicago. Plotting the change in crime-rate over time will help us to identify specific crimes or specific locations that we should investigate further. We will be using various python libraries such as folium and seaborn to create useful visualizations that we can use to spot trends in the data. Folium will help us to easily heat map the data, allowing us to see which neighborhoods are worth looking into. This will help us to better understand how a particular neighborhood can get progressively worse over time.

4 DATA SET:

The main dataset comes from Chicago's online data catalog. The dataset lists roughly 6.54 million crime reports from 2001 to present. The columns that are most important to our project include the date when the crime occured, the Illinois Uniform Crime Reporting code, the ID of the community, a short description of the crime, a attribute indicating whether an arrest happened, the police district, and the geographical coordinates. There are other columns in the dataset, such as case number of the incident, the FBI code, or the ward ID, but we will unlikely utilize them. This dataset is about 1.5 Gigabytes on disk, which will pose some challenges for data manipulation. From initial data analysis, it looks like some of the data will need to be preprocessed to clean up invalid dates and missing data.

In addition to the main dataset of crime reports in Chicago, there are several other interesting datasets available through the city's data catalog. One of the datasets available describes various socioeconomic indicators of the different communities in Chicago. For example, it lists the percent of housing that is overcrowded, percent of households living below the federal poverty line, percent of people aged 16 and older that are unemployed, per capita income, and percent of persons over the age of 25 without a high school diploma. Given these community based indicators, it might be possible to find a correlation between these values and the rates of crime in the region. Other interesting datasets available on the catalog provide information on affordable housing and grocery stores in Chicago's communities. The catalog also publishes public health information such as the rates of elevated blood lead levels in children aged 0-6 years old as well as data on births and pregnancy care by community area. Like the socioeconomic indicators, these community statistics could provide less obvious links to crime rates.

5 EVALUATION METHODS:

Firstly, we hope to arrive at comparable results to previous studies of Chicago's crime. Secondly, we want to go further than previous studies and find novel trends between crime and other social and public factors. This knowledge could potentially influence Chicago's crime reduction strategies. For example, if we found that affordable housing was correlated with lower rates of crime that would be valuable knowledge for the Chicago city planners.

6 TOOLS:

6.1 Python:

The majority of our programming will be done in python. Python has a lot of useful tools that will help us work with this large data set. Some of the libraries we will be using in Python are Pandas, Numpy, Matplotlib, Folium, and Scipy.

6.2 Jupyter NoteBooks:

Jupyter notebooks are easy to use and great for exploring ideas quickly with python. Most of our work will likely involve jupyter notebooks in some way.

6.3 Pandas:

We will be using pandas to handle most of the data manipulation such as cleaning and structurization. Pandas will allow us to manipulate the data so that we can create better, more interesting visualizations.

6.4 Numpy:

It will be used for scientific computation and working with arrays.

6.5 Folium:

Used for mapping instances in the data set. We will use folium to heatmap different types of crime. This will help us to visualize the data and spot potential patterns to investigate.

6.6 Matplotlib:

This package will help us visualize the data after being processed. It will help us generate plots, histograms, bar charts, scatterplots, etc., with just few lines of code.

6.7 Scipy:

This library will be used to calculate statistics of our data. It will facilitate to calculate averages, means, medians, modes, z-scores and more statistical values.

6.8 RapidMiner:

Aside from coding in python, we will be using RapidMiner software to get predictive analytics and statistical modeling.

7 MILESTONES:

Milestone 1: Have the data preprocessed - March 9th.
Milestone 2: Create visualizations - March 16th. Augment crime data with another data set or sets - March 16th.

Milestone 3: Create heat maps based upon districts - March 23rd.
Milestone 4: Dive deep into the ’Weird’ data - April 6th.
Milestone 5: Implement AI eg. (clustering, neural network, knn or random forest) - April 20th.

8 PROGRESS REPORT:

The main dataset used in this project is comprised of over 6.5 million rows of crime records spanning 2001 to February 2018. This obviously poses a challenge to load into a laptop computer’s memory. In order to make the dataset easier to manipulate, it was broken up into 18 new datasets grouped by year. The records in these datasets describe how many total crimes, index crimes, and crimes by type that were committed each day in each Chicago community area. Aggregating the data by day and community area ID reduced the entire dataset to around 44,000 records, which made it much easier to handle. Pandas, the Python library, was used heavily for data manipulation in this project.

Once this was done, these yearly datasets were used to plot the amount of crime committed on each day of the year(Figure 1). Additionally, the crime by type (such as ”battery”, ”theft”, etc.) was plotted for each year(Figure 2). Examining these plots, it seems that the city of Chicago did not record all crime records from the beginning of 2001 until around March of 2002. This is apparent in the drastic change in the plot of 2002 crime as crime jumped from at maximum 150 crimes in one day before March to around 1300 per day in April. Another feature that can be observed from these graphs is that the rate in total crime and index crime follows a fairly consistent pattern. The shape of the scatter plot is slightly bell-shaped where the number of crimes is highest roughly May through July. Interestingly, the number of of total crimes spikes in early January especially near New Year’s Day.

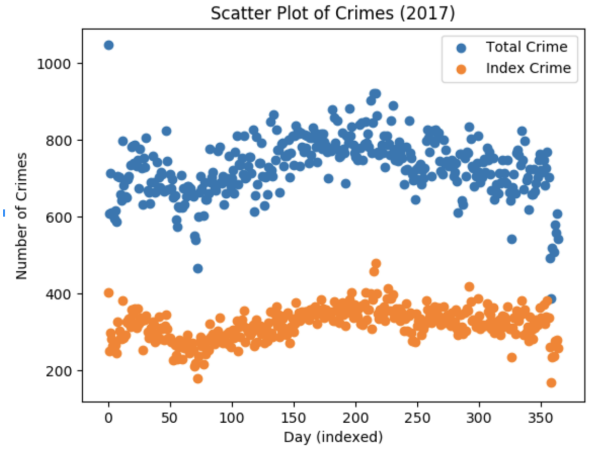

Figure 1: ***Scatter Plot of Crimes***

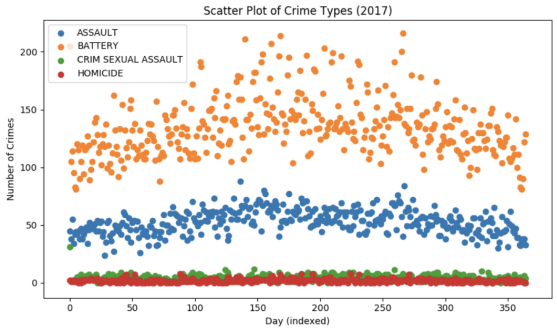

Figure 2: ***Scatter Plot of Crime Types***

The yearly datasets were combined into a single dataframe for crime across all 18 years studied. From this graph (Figure 3) it was easy to see that overall the total number of crimes in Chicago steadily declined from 2003 to present-day 2018. The years 2001 and 2002 were excluded from this plot because of the issue of missing data for that time period mentioned above. In addition, a simple linear regression was run on the dataframe using Scipy and it too revealed a declining rate (a negative slope). Numpy’s polynomial regression method was also run on the data. It performed better than the linear method because it could fit better to the month to month fluctuations. This method also showed a decreasing amount of total crime.

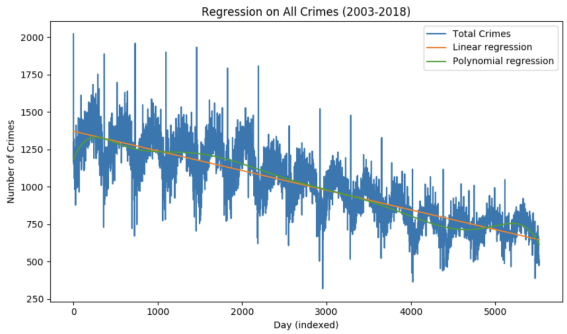

Figure 3: ***Regression on all Crimes***

A third task was finding the outliers in the crime data. This, like the plots above, was done year to year and then overall the entire time range. A subset of crime types were examined including: assault, battery, criminal sexual assault, homicide, stalking, and prostitution. These crimes were chosen out of pure curiosity, but further analysis should investigate all crime types. For each year, the program found the mean and standard deviation in the number of crimes committed of that type. Outliers were found by looking for days in which the number of crimes was above or below a certain number of deviations. In this case, we used three standard deviations and only looked at the days where the crimes exceeded this amount (not the days where it was less than it). The program would find the days in the year that matched this criteria and logged it to a file. In addition, the number of total crimes of that type was summed up and plotted across the year. Several

interesting trends were seen from this data. Firstly, many of the outliers are around the middle of year which reflects the overall trend described above. Secondly, crimes of theft are particularly high on the first of the month. Secondly, homicides are higher during the early June to late September period. Thirdly, cases of battery and criminal sexual assault are particularly high around New Year's. Further analysis of this outlier data will break up the data by community area to find which regions are notably outside the normal rate of crime.

One of the goals described in the project proposal was to incorporate multiple datasets in this project. These datasets would describe additional information about Chicago including census data, social data, and public health. Four datasets from the city of Chicago's online data catalog were added to yearly dataframe. The first dataset listed the number of affordable housing units per community area in Chicago. The second dataset provided information on elevated blood level levels in children ages 0-6 between the years of 1999-2013. The third dataset gave data on rates of prenatal care by community region in Chicago from 1999 to 2009. The fourth dataset listed static socioeconomic figures for each community region, including information about the overcrowded housing, households below the poverty line, the percent unemployed or without a high school diploma, and per capita income. The crime dataframe was then grouped by community area in order to get the total number of crimes committed in a community per year. Using these yearly crime totals and social metrics from the other datasets, a correlation matrix was created using Pandas. From this matrix it appears that there is a slightly positive correlation in the amount of crime and the rate where no prenatal care was given. Similarly, there were slight positive correlations with features of elevated blood lead levels, percent below poverty line, and percent unemployed. However, this matrix needs to be analyzed very warily for a couple reasons. Firstly, the social data used was either static or did not span the same years of the crime dataset (2001-2018). In order to handle this, the missing values were filled with the average value in that column. This of course skews the data because it does not take into some of the community changes not described in the time range. For example, the per capita income of each of Chicago's communities probably did not remain static from 2001 to 2018. Secondly, the correlation needs to take into more granular data about each community's crime statistics instead of just overall crime rates. Despite this, the matrix displays other interesting data such as the correlation between prenatal care rates and poverty or unemployment rates and blood lead levels.

In addition to the work described above other tasks were worked on. One of the tasks attempted was to find a correlation between the day of the year, the community region, and the number of a type of crime committed. However, this correlation matrix showed only weak support the way it was constructed. Alternative implementations should be explored to discover the correlations between community, time of the year, and crimes rates. Another task attempted was to predict the community region given crime record that lists the type of crime, the location type, whether a an arrest was performed, whether it was a domestic crime, and the year it was committed. ScikitLearn was used to train a multiclass random forest classifier to identify the community region. This process was run on all 6.5 million rows of the dataset and after 11 hours of running it showed no signs of stopping. Future investigation should try random sampling to get a smaller subset of the entire dataset for training. A third task, which is currently in development, is to predict the number of crimes of one type given the number of crimes of another type. For example, given the number of crimes of narcotics, robbery, theft, burglary, stalking, and assault, can you predict the number of sexual assault crimes that will occur on that day and in a specific community?

In an attempt to better understand the data set we created some visualizations in order to gain some additional insight. Some of the most useful visualizations that we created were heat-maps generated using the Python library Folium (Figure 4). Fortunately, Folium works well with Pandas which allowed us to easily create heat-maps of different types of crimes.

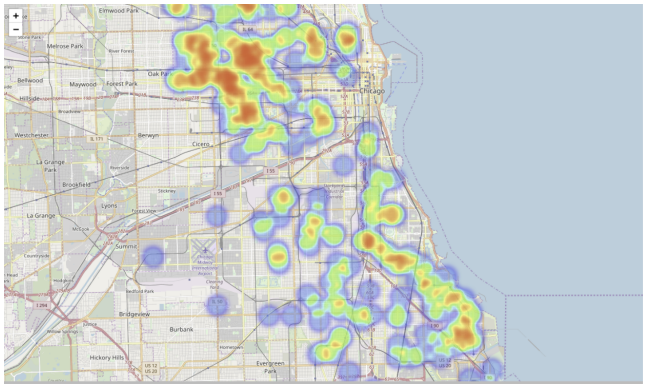

Figure 4: ***Heat-map of Crimes in Chicago***

One particularly interesting trend that we noticed when we were comparing different heat-maps was the drastic change in the number of prostitution arrests between the years 2001-2017(Figure 5 - Figure 6).

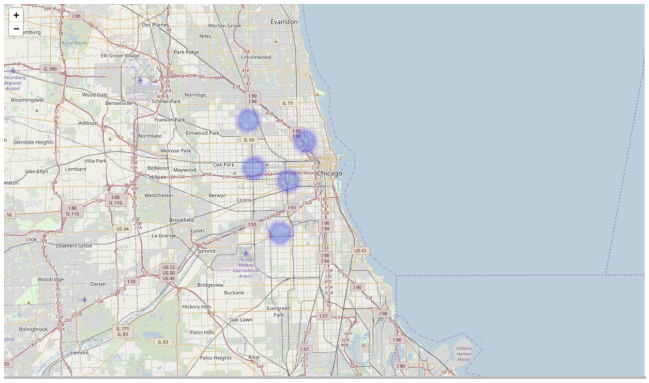

Figure 5: ***Heat-map of Crimes in Chicago, 2001***

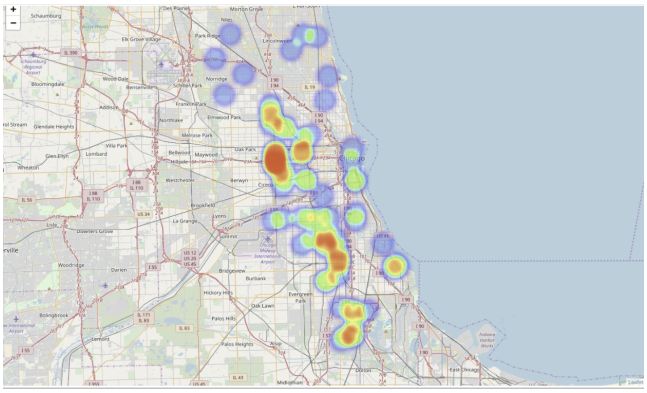

Figure 6: ***Heat-map of Crimes in Chicago, 2017***

Another trend that we spotted while doing visualizations was the drastic change in crime rate in the 100xx block of W O’Hare Street. Which, in the early 2000’s had one of the worst crime rates in Chicago. But, after a few years, the crime rate in that area decreased drastically. We were wondering if there was a simple reason for this sudden decline. We did some research and discovered that the City of Chicago had invested a significant amount of money in efforts to revitalize the area. Who would have thought that there was such a simple way to decrease the crime rate in an area, invest some time and money in making it nicer. We will likely spend some more time exploring similar areas to see if there are any noticeable changes in the type of crimes or volume of crimes in a particular area.

9 WORKS CITED:

Mangipudi, Vivek. ANALYSIS OF CRIMES IN CHICAGO 2001 - 2017. 28 July 2017, rstudio-pubs-static.s3.amazonaws.com/294927_b60231 8d06b74e4cb2e6be336522e94e.html

Papachristos, Andrew V. fi48 YEARS OF CRIME IN CHICAGO: A Descriptive Analysis of Serious Crime Trends from 1965 to 2013.fi Yale ISPS, vol. 13, no. 023, 9 Dec. 2013, pp. 1fi?!20., isps.yale.edu/sites/default/ files/publication/2013/12/48yearsofcrime_final_ispsworkingpaper023. pdf

Laughlin, Greg. fiCrime Over Time: Visualizing Crime Data in Chicago.fi Socrata, 3 June 2014, socrata.com/blog/crime-time-visualizing-crime-data-chicago/.